POTION

Optimization of time-frequency transforms for audio coding using a perceptive measure of distortion and a sparsity constraint

Ichrak Toumi & Olivier Derrien - LMA

FLAME Meeting, November 2014

laboratoire de mécanique et d'acoustique

AGENCE NATIONALE DE LA RECHERCHE ANR

cnrs dépasser les frontières

Outline

- 1. Compression of audio signal and time-frequency transforms
- 2. Writing compression as a standard optimization problem
- 3. Choosing suitable time-frequency transforms
- 4. Preliminary results with union of MDCTs
- 5. Perspectives

1. Compression of audio signals and t-f transforms

Definition of audio coding

- Minimize the amount of information to be stored/transmitted for near-perfect audio quality
- or maximize the audio quality for a given amount of information

Link with time-frequency transforms

- State-of-the art codec (MP3, AAC) rely on invertible time-frequency transforms (PQMF-Banks, MDCT)
- because audition can be efficiently modelized in the TF domain

Sparsity in audio coding

- Reducing the amount of information is achieved by
  - Setting some coefficients to zero
  - Re-quantize non-zero coefficients
- Example: AAC @ 128 kbps
  - Stored/transmitted information: 10% of the original
  - Non-zero coefficients: 30% of the original
  - Remaining 20%: re-quantization
- A sparse representation is desirable for audio coding
- Target sparsity value : 30% non-zero coefficients or less

2. Writing compression as a standard optimization problem

- Consider discrete-time, N samples long, real-valued signals: $\mathbf{x} \in \mathbb{R}^N$

The coding transform

- Consider a time-frequency transform characterized by
- An analysis dictionary: $\mathbf{A} = \left\{\phi_1^H \cdots \phi_M^H\right\}$, $\phi_m \in \mathbb{C}^N$
- A synthesis dictionary: $\mathbf{S}^T = \left\{\psi_1^T \cdots \psi_M^T\right\}$, $\psi_m \in \mathbb{C}^N$
- The analysis operator is: $\mathbf{y} = \mathbf{x}\,\mathbf{A} \quad \Leftrightarrow \quad y_m =< \mathbf{x}, \phi_m > \ \forall m$
- The synthesis operator is: $\hat{\mathbf{x}} = \mathbf{y}\,\mathbf{S} \quad \Leftrightarrow \quad \hat{\mathbf{x}} = \sum_m y_m\, \psi_m$
- Perfect reconstruction $\Leftrightarrow$ $\mathbf{A}\,\mathbf{S} = \mathbf{I}_N$, which implies $M \geq N$

2. Writing compression as a standard optimization problem

The perceptual transform

- A relevant measure for perceived distortion can be computed using a perceptual time-frequency transform of size $Q \geq M$
- The analysis dictionary is: $\mathbf{P} = \left\{\mathbf{p}_1^H \cdots \mathbf{p}_Q^H\right\}$, $\mathbf{p}_q \in \mathbb{C}^Q$
- There is no need for a synthesis dictionary
- We assume that perceptual weights $\mu_q > 0$ associated to each vector $\mathbf{p}_q$ can be computed using an audition model

The perceptual distortion measure

$$D_p = \| \left(\mathbf{x} - \hat{\mathbf{x}}\right) \mathbf{P}\Delta_\mu \|^2$$

with $\Delta_\mu = \text{diag}\left(\mu_1, \cdots, \mu_Q\right) \Rightarrow D_p =$ weighted L2 norm of the error

2. Writing compression as a standard optimization problem

Re-writing the perceptual distortion measure

$$D_p = \| (\mathbf{x}\,\mathbf{P} - \mathbf{y}\,\mathbf{SP})\,\Delta_\mu \|^2$$

Formulating the coding problem

- Find $\mathbf{y}$ that minimizes D_p
- If we consider the quantization of y_m, $\mathbf{y}$ is searched only in a finite subset of $\mathbb{R}^K$. *That will not be considered for the moment*
- This is a weighted-L2 optimization problem of the form:

$$\text{Argmin}_{\mathbf{y}} \left[\| (\mathbf{g} - \mathbf{y}\,\mathbf{K})\,\Delta_\mu \|^2 \right]$$

- where $\mathbf{K} = \mathbf{SP}$ is called the mixture matrix (size $M \times Q$)

2. Writing compression as a standard optimization problem

Finding solutions to the coding problem

- The existence of solutions mainly depends on the properties of $\mathbf{K}$
- If $\text{rk}(\mathbf{K}) = M$, the solution is unique: $\tilde{\mathbf{y}} = \mathbf{g}\,\mathbf{K}^{\dagger}$
- Otherwise, there is an infinite set of equivalent solutions
- For selecting "the best" solution, or when $\mathbf{K}$ is badly conditioned, one usually add a regularization term that promotes sparsity:

$$\text{Argmin}_{\mathbf{y}} \left[\| \left(\mathbf{g} - \mathbf{y}\,\mathbf{K}\right) \Delta_{\mu} \|^2 + \lambda \| \mathbf{y} \|^p \right]$$

- Finding a sparse solution is especially desirable in audio coding

3. Choosing suitable time-frequency transforms

Choosing a perceptual transform: $\mathbf{P}$

- Constrained by the existence of an earing model to compute μ_q
- DFT or MDCT: work with standard MPEG hearing models
- Constant-Q or ERBLett: more sophisticated models available

Choosing a coding transform: $\mathbf{A}$ and $\mathbf{S}$

- Audio signals should naturally have sparse representations in the transform domain
- Perfect reconstruction is not necessary
- The choice shall depend on the rank of $\mathbf{K} = \mathbf{SP}$
- If $\text{rk}(\mathbf{K}) \ll M$, there are many local minima and the practical solution strongly depends on the initialization
- A good choice corresponds to $\text{rk}(\mathbf{K}) \simeq M$

3. Choosing suitable time-frequency transforms

Solutions that work

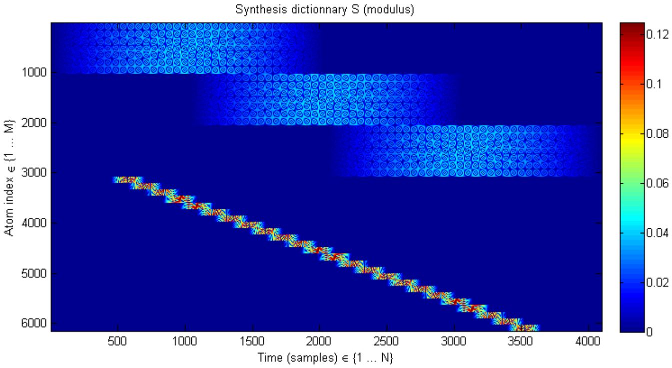
- $\mathbf{A} = \mathbf{P}$ is a single MDCT
  Then $\mathbf{S} = \mathbf{A}^T$ and $\mathbf{K} = \mathbf{I}_M \Rightarrow$ the problem is diagonal
  This is a trivial case: the solution is obtained by thresholding $\mathbf{g}$
- $\mathbf{A} = \mathbf{P}$ is union of MDCTs with different sizes
  Then $\mathbf{S} = \mathbf{A}^T$ and $\mathbf{AS} \neq \mathbf{I}_N \Leftrightarrow$ no perfect reconstruction
  $\mathbf{K} \neq \mathbf{I}_M$ and $\mathrm{rk}(\mathbf{K}) < M \Leftrightarrow$ many local minima
  But $\mathbf{K}$ is a very sparse matrix: when thresholding very small values to zero we get $\mathrm{rk}(\mathbf{K}) = M$

Solutions that does not work (for the moment)

- $\mathbf{A}$ is a MDCT and $\mathbf{P}$ is an ERBLett
  $\mathrm{rk}(\mathbf{K}) \ll M$ and the problem can not be regularized properly
- But things seem to get better with the *real part* of an ERBLett

Analysis/synthesis matrix

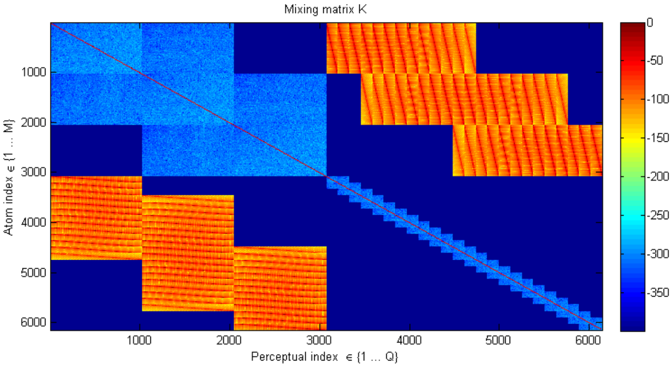
- We choose the union of 2 MDCTs: 1024 bands and 128 bands
- idem AAC, but here both MDCTs can be used simultaneously
- For plots, we choose $N = 4096 \Rightarrow M = 6144$



4. Preliminary results with union of MDCTs

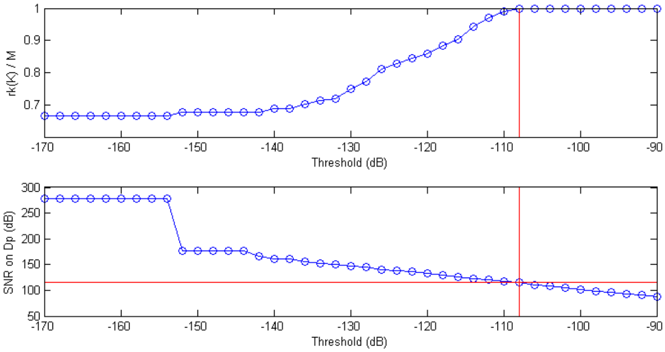
Perceptual matrix and mixture matrix

- We assume $\mathbf{P} = \mathbf{A} = \mathbf{S}^T \Rightarrow Q = M$
- Then $\mathbf{K} = \mathbf{S}\,\mathbf{S}^T \Rightarrow K(m,q) = <\phi_m, \phi_q>$
- $\text{rk}\,(\mathbf{K}) = N = 4096 < M = 6144$



Thresholding the mixture matrix

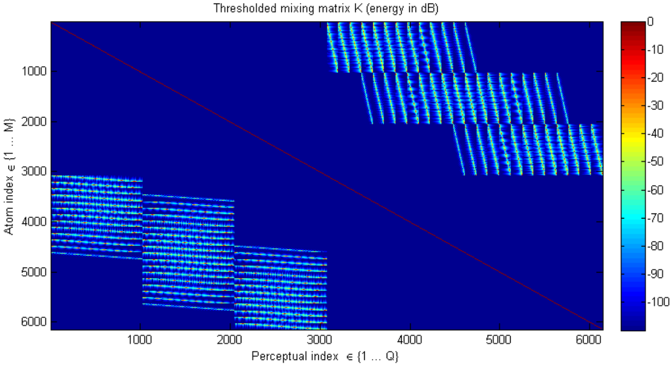
- We set a threshold T so that $K(m,q) \mapsto 0$ if $K(m,q) < T$
- This implies an error on the estimation of the distortion D_p
- $T = -108$ dB $\Rightarrow SNR = 110$ dB (near perfect) and $\mathrm{rk}\left(\mathbf{K}\right) = M$



4. Preliminary results with union of MDCTs

Thresholded mixture matrix

- $T = -108$ dB
- $\mathrm{rk}\,(\mathbf{K}) = M \Rightarrow$ there is a unique solution to the optimization problem, i.e. the approximation of D_p is convex



4. Preliminary results with union of MDCTs

Implementations details

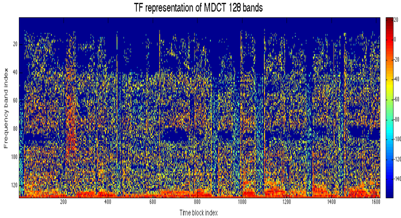
- The perceptual weights μ_q are computed for both resolutions (1024 and 128 bands) with the MPEG #2 hearing model
- The target $\mathbf{g} = \mathbf{x}\,\mathbf{P}$ is computed using a standard MDCT implementation
- The thresholded mixture matrix is stored as a sparse matrix
- The signal is divided in macro-blocks, and the optimization is performed independently on each macro-block
- No redundancy is added when macro-blocks overlap
- The sparsity level is set by the regularization constant λ

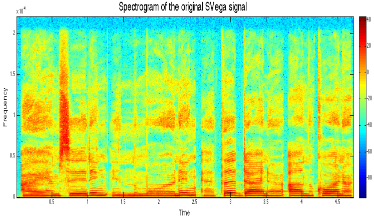
4. Preliminary results with union of MDCTs

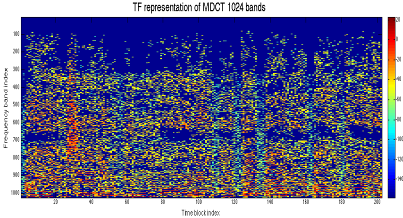
Sparsity rate = 43 %

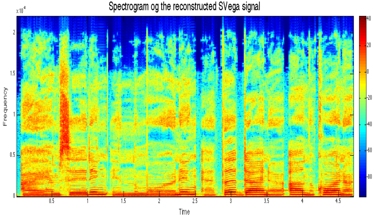
SVega original signal

SVega Reconstructed signal





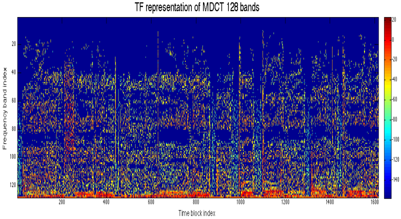


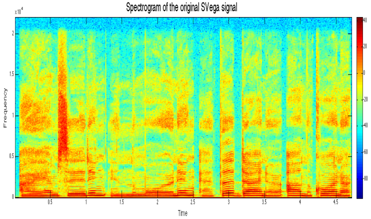


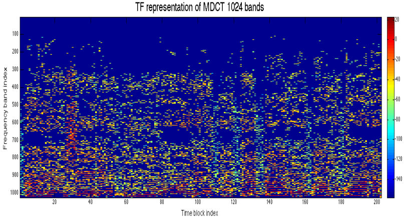
4. Preliminary results with union of MDCTs

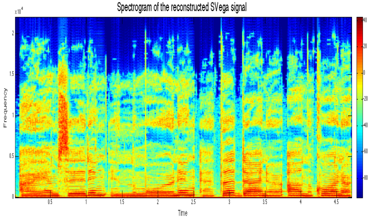
Sparsity rate = 66 %

SVega original signal   SVega Reconstructed signal







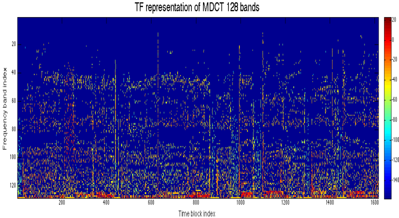


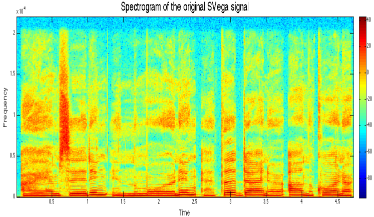
4. Preliminary results with union of MDCTs

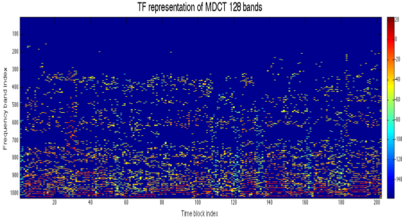
Sparsity rate = 83 %

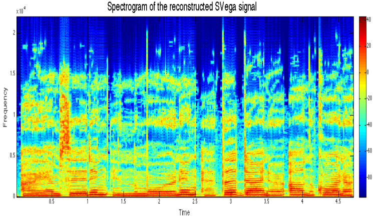
SVega original signal

SVega Reconstructed signal









5. Perspectives

- Try different time-frequency transforms for **S** and **P** in order to find the couple which offers the best tradeoff between perceived audio quality and sparsity rate
- Try a more sophisticated perceptive model, different from the MPEG #2
- Include the quantization step in the optimization algorithm